基于人工智能技术治理网络暴力的探析

孙 仕

（广东财经大学公共管理学院，广东 佛山 528100）

摘 要：【目的】随着信息技术的高速发展，使得人类社会的交流方式、娱乐方式及获取信息的方式得到了多样化的发展，同时也催生出各类问题和挑战。人工智能技术合理有效规范地运用是现代社会问题治理的重要手段，关系到精准化社会治理的有效实施，文章旨在探析引入人工智能技术治理网络暴力；【方法】文章厘清网络暴力的概念、成因和危害，结合人工智能的相关概述，分析引入人工智能技术精准化治理网络暴力的可行性和具体应用案例；【结果】人工智能技术可以有效拦截和屏蔽网络暴力语言、图片和视频，并提供有效的干预，从而减少网暴行为的发生；【结论】人工智能技术可以极大程度阻止网络暴力的发生，并且人工智能技术治理网络暴力的应用可以推动公共治理方式的创新。

关键词：人工智能；网络暴力；精准化治理；公共治理创新；数字治理 **中图分类号：**G223 **文献标识码：**A

文章编号：1671-0134（2023）01-064-06 **DOI：**10.19483/j.cnki.11-4653/n.2023.01.011

本文著录格式：孙仕 . 基于人工智能技术治理网络暴力的探析 [J]. 中国传媒科技，2023（01）：64-68，87.

1. 问题的提出

当代以互联网为代表的信息技术的高速发展，让人类社会生活变得更加便利化和多样化。同时也催生了与以往完全不同的社会问题，各类社会性问题频发且呈现剧烈化和扩散化的趋势，公共治理压力也随之增加，亟需相关治理理论的创新和治理能力的提高，也更加促成了人们对利用信息技术手段来解决公共问题的创新和思考。基于社交网络快速发展下产生的网络暴力就是当下最典型的社会现象，它带来的实时性、集中性和直接性的伤害让许多人深受其害。也正因如此，探索出更多解决网络暴力的方法就势在必行。

2. 人工智能概述

通过自然的进化与发展而产生的智能，被称为自然智能。人类智能是目前已知的自然智能中最复杂且最高级的智能。通过人类智能而间接制造和发展出来的智能，被称为人工智能或机器智能。人工智能来源于自然智能，特别是人类智能。所以，人工智能研究的首要任务就是认识和理解自然智能，创造人工智能的机器和应用，增强并服务人类智能的开发。[1] 人工智能的概念最早由 John McCarthy 于 1955 年在达特茅斯会议上正式提出。人工智能（AI：Artificial Intelligence），根据英文释义可以理解为“人造的智能”。人工智能被划分为了三个阶段，即弱人工智能阶段、强人工智能阶段和超级人工智能阶段。弱人工智能是利用智能化技术改善经济和社会发展所需的一些技术条件和功能；强人工智能是十分接近于人类智能的阶段，目前普遍认为这个阶段要持续发展到 21 世纪中叶后才能真正实现；超级人工智能是脑科学和类脑智能有了突破发展后而创造的一个超越人类智能的超强智能系统。当下的人工智能发展还处于弱人工智能阶段，研究领域主要是语音识别、图像识别、自然语言处理、智能机器人、专家系统和自动驾驶等方面，延伸出的科技领域已经涉及人类社会的多个方面。[2]

3. 网络暴力概述

本文通过我国最大的文献数据库——中国知网（CNKI），检索了该概念的学科研究状况，主要涉及新闻传播学、政治学、法学及社会学等多个学科，目前已经初步形成规模（见图 1）。其中不少学者均对网络暴力进行了概念描述，如姜方炳（2011 年）将网络暴力定义为“网络技术风险与网下社会风险经由网络行为主体的交互行动而发生交叠，继而可能致使当事人的名誉权、隐私权等人格权益受损的一系列网络失范行为”。[3] 陈代波（2013 年）认为网络暴力是“网民对当事人或者组织实施的以制造心理压力为手段，以迫使当事人或者组织屈服的网络攻击性行为的总称；因这种行为带有明显的强制性特征，与现实中的暴力类似，故而称为网络暴力”。[4] 路芳（2010 年）将网络暴力界定为“网络暴力是一种作为行为施方的网络行为主体以其隐蔽性、强制性、极端性和侵犯性的网络行为给行为受方造成实质性伤害的网络行为失范”。

[5] 他们都在强调网络暴力性质和特征的基础上进行相应的概念厘定。于此，本文将网络暴力总结为：网络行为主体通过网络社交渠道以极端化和强制化手段，对当事人造成心理上实质性伤害的类暴力行为的总称。最普遍的网络暴力方式包括网络造谣、网络诽谤、人肉搜索、信息曝光、道德谴责、网络骚扰、网络恐吓等行为。

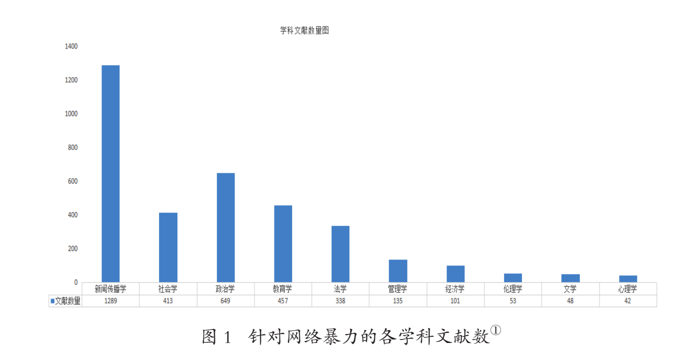


图 1　针对网络暴力的各学科文献数①

网络暴力的成因来源于多方面的因素。首先，互联网的开放性和匿名性，导致网络行为主体人（包括网络行为个人、网络行为商业团体等）在原有的社会事件中忽略客观事实，为了增加关注度或是为了商业盈利而刻意加入主观的歪曲见解来制造矛盾，经过多方的转发和歪曲叠加后，最终让阅读者和观看者得到与客观事实完全不符的信息，这种网络信息失真会激怒部分极端主义者而发起网络暴力的行为。

其次，网民低龄化和整体网民受教育程度偏低也是重要因素。根据中国网信办（CNNIC）发布的《2021年全国未成年人互联网使用情况研究报告》的统计数据显示："2021 年未成年网民达到 1.91 亿，未成年人互联网普及率达 96.8%，较 2020 年（94.9%）提升 1.9 个百分点"。②第 50 次《中国互联网络发展状况统计报告》显示："截至 2022 年 6 月，我国网民规模达 10.51 亿，较 2021 年 12 月新增网民 1919 万，互联网普及率达 74.4%"③；第 47 次《中国互联网络发展状况统计报告》中网民学历结构的数据显示："小学、初中、高中 / 中专 / 技校学历的网民占比分别为 19.3%、40.3% 和 20.6%；受过大学专科、大学本科及以上教育的网民占比分别仅为 10.5% 和 9.3%"。④通过以上数据可以发现，我国网民以中等教育水平的群体为主，网民不光低龄化而且整体受教育程度不高。

再者，随着互联网数字经济在我国的爆发式增长，各种基于互联网发展起来的科技公司得到了进一步的发展，比如腾讯、百度、新浪、优酷、字节跳动等，均开发出了相应的社交 App、视频软件及论坛，从 Web1.0 时代转向 Web2.0 时代，为网络行为主体提供了动态化、参与式、可读可写以及便捷化的互动方式，但也为网络暴力孕育了温床。[6] 由于网暴者本身也是这类社交平台的"客户"，在盈利需求的驱使下，会疏于自我监管机制建设，最终助长网络暴力行为的不良风气。

最后，网络暴力法制建设缺失也是不容忽视的因素。法律是公民行为规范的基本准则，针对性的立法必然是控制网络暴力的有效手段，但是我国仍然将网络暴力相关的行为归类于民事侵权的范畴，尚且未形成网络暴力专项立法。一旦发生网络暴力行为，目前仅能依据部分条例和规定进行维权。另外，由于互联网的匿名性、网络暴力行为主体的多元性、网暴受害者受害程度难以估量性等多重因素，也导致受害者容易因为取证艰难而陷入维权困境。

4. 人工智能精准化治理网络暴力的可行性

4.1　治理网络暴力的迫切需求

4.1.1　传媒环境的净化

"沉默的螺旋"理论是指每个人生来就是害怕孤独的，所以在面对具有争议性的话题时，他们会试图判断自己的意见是否属于大多数，然后判断公共舆论是否会朝着赞同他们意见的方向改变，一旦他们觉得自己的意见占劣势，在"担心被孤立"心理作用下就会倾向于对该话题保持沉默。因此，舆论占优势的一方便会更加得势，舆论处于劣势的一方便会无限地沉默下去，形成了一个"螺旋式"的传播过程。[7] 网络是个开放的世界，任何人都可以表达自己的观点和看法，但是网络暴力加速了"沉默螺旋"的发展，阻碍了少数人发表观点和看法的机会，破坏了开放、包容、自由的网络传媒环境。其次，传媒在社会治理中起到监督作用，但是网络暴力让"监督"超越了"边界"，以致于让不法个人和机构、商业团体为了利益把监督变成了"舆论造势"。这种"舆论造势"再加上自媒体时代的"蹭热度"行为，就把良性的媒体监督变成了"开局一张图、内容全靠编"的网络行为，不光损害大众的利益，也严重降低大众传媒的公信力。网络暴力导致的传媒环境恶化问题，亟待解决。

4.1.2　公众权益的保护

网络暴力严重损害公众利益。网络暴力行为主体者的侮辱、造谣、谩骂及其他不正当网络行为会给当

事者及其家人在精神上造成创伤而影响正常生活的开展。近年来，众多名人、学者及普通人都因为遭受网络暴力而患上抑郁症，甚至还会选择轻生自杀。比如刘学洲事件⑤、王力宏离婚事件⑥、电视剧虐猫事件⑦等。同时，网络暴力扭曲大众道德价值观。网络暴力事件最常见的例子就是站在个人的道德标准上对当事者进行道德审判，呈现出道德判断的偏激和片面。由于互联网的开放性、匿名性及其相应的法制缺失，让这种网络行为极易超出正常的边界，在得到愤世嫉俗的审判快感后，网络暴力行为主体者对当事者精神上摧残的背后也无需对此承担任何责任，甚至还能彰显出道德的高尚和伟大，这种肆无忌惮的网络行为方式扭曲了大众的道德判断标准和公共道德底线。[8]

4.1.3 企业和国家的稳定发展

网络暴力阻碍互联网行业的良性发展。网络暴力作为互联网时代高速发展衍生出的一个“毒瘤”，它不仅会严重影响大众的网络冲浪体验，同时也会给网络运营商和网络社交平台带来相应的损失，比如品牌形象受损、核心用户流失等，甚至可能导致企业的商业投资撤回、破产和倒闭。而且，网络暴力影响社会的稳定和国家安全。网络行为主体者借助网络平台散播的攻击性和煽动性的言论、敏感话题相关的图片及视频，经过互联网的扩散和发酵后会出现在网络的各个角落，这种无限拓展性和不可控性，最终会形成“网络舆论一边倒”现象。相比于传统媒体舆论，网络舆论具有多元复杂性、即时互动性和隐蔽性等特点。如果网络暴力形成的类似于“民主”“人权”等话题的网络舆论，也极易被西方或敌对势力所利用，策划或煽动反动活动，攻击我国的政治制度，以致影响我国的社会稳定和国家安全。[9]

4.2 公众对人工智能的高接受度

现有的人工智能应用对人们的生活方式、学习方式、娱乐方式都产生了巨大的影响。不仅大幅提高了社会资源的使用效率，也是让生产方式更加趋于人性化。据德勒的一项民意调查显示，“68% 的中国用户对人工智能技术持积极态度，用户认为人工智能技术将对社会发展、教育、医疗水平、环保和社会公平都有积极作用”。⑧利用人工智能技术解决网络暴力问题目前已经成为学术界和社会日益关注的热点，强大的治理需求促使了人工智能治理网络暴力的进一步实施和开展。

4.3 国家政策的大力支持

为促进人工智能技术的发展，国家发展改革委、科技部、工业和信息化部、中央网信办联合制定了《“互联网 +”人工智能三年行动实施方案》（2016）、国务院印发了《新一代人工智能发展规划》的通知（2017）、工业和信息化部印发了《促进新一代人工智能产业发展三年行动计划（2018–2020 年）》（2017），2019 年政府工作报告中更是将人工智能升级为“智能 +”。国家政策对人工智能领域的引导和鼓励，促进我国从“互联网 +”时代向“人工智能 +”时代的积极转变。另一方面，为更好地防治网络暴力问题，切实地保障网民自身的合法权益，以中央网信办为主的相关部门也积极开展了打击网络暴力和改善网络环境的专项行为，比如 2022 年 4 月部署和开展了针对影响力较大的 18 家网络平台的“清朗・网络暴力专项治理行动”和 2022 年 8 月举办的中国网络文明大会等。

4.4 人工智能支撑技术的高速发展

“数据”“算力”和“算法”是人工智能最重要的 3 个要素，它们之间相互促进和相互支撑，最终促成了人工智能技术的应用和价值创造。自党的十九大以来，数字中国建设取得显著成效，基本建成全球规模最大且技术领先的网络基础设施。“截至 2021 年底，我国已建成 142.5 万个 5G 基站，总量占全球 60% 以上，5G 用户数达到 3.55 亿户”。⑨完善的网络基础设施建设促进了我国互联网产业的快速发展。网络暴力产生于互联网，而互联网中用户每一次交互过程都会留下后台痕迹，随着交互记录不断在后台系统积累，最终就会形成庞大的数据库，数据库存储的数据为人工智能的训练提供了先决条件，技能获取后的人工智能可以根据不同场景的数据得到相应的智能模型。算力，即计算机处理数据的能力，也是人工智能突破发展的决定因素。根据 2022 年中国算力大会公布的数据显示，“截至 2022 年 6 月底，我国在用数据中心机架总规模超过 590 万标准机架，服务器规模约 2000 万台，算力总规模超过 150EFlops。与此同时，算力产业链条持续完善，包括算力基础设施、算力平台、算力服务等在内的、具有国际竞争力的算力产业生态初步形成，一批具有示范效应的算力平台、新型数据中心以及产业基地相继落地”。⑩算法模型、数据技术和算力产业的蓬勃发展也为人工智能技的发展打下了夯实的基础，出现了如百度 AI、华为、商汤及阿里巴巴等涉及人工

智能技术的公司和品牌。

4.5 精准化治理网络暴力的实践应用

4.5.1 敏感词、敏感音频及敏感图像和视频

对以自然语言描述为表现形式的网络暴力行为，目前最常用的人工智能技术是“敏感词过滤”。建立一个“敏感词文字描述库”作为检测自然语言描述的依据，通过后台的智能识别和智能过滤，将过滤的敏感词直接屏蔽、替换或阻止发布，减少网络暴力敏感词的文字传播，从而减少语言攻击和降低语言暴力伤害。[10] 由于音频数据主要以波形型号为主，所以对以音频为表现形式的网络暴力行为，主要是利用波形特征对网络暴力特征进行提取、识别和分类。对于以图像和视频为表现形式的网络暴力行为，一般需要提取图像和视频的特征，针对图像和视频中目标运动的动作、轨迹和姿态等特征，采用多分类算法模型对不同类别的网络暴力行为进行识别后，定位相关语义特征的图像和视频，最后实现对带有血腥、恐怖组织、枪支器械、色情、垃圾营销广告、水印等诸如此类网络暴力行为的图像和视频的过滤和删除。目前，国内各大网络平台均有针对敏感词、图像和视频及音频的应用，如新浪微博、微信及抖音等平台。另外，诸如百度 AI 智能云、网易易盾和数美之类的专业第三方内容审核服务也日趋成熟（见图 2）。



图 2 百度 AI 审核服务、数美智能检测服务

4.5.2 人工智能提示和服务

人工智能“语言提示”和“用户自主拉黑”功能，是利用人工智能技术，在暴力性评论语言发表之前，后台会自动向评论者发出提醒，让评论者更改语言来保持友善的社交环境。用户的个人账户下如果经常受到来自某个或多个人的带有辱骂、攻击或暴力的评论，该用户可以设置“拉黑”功能来防止此类评论的继续发表。这两项功能均可以帮助预防网络暴力。目前国内外的众多社交平台均有上线诸如此类的人工智能服务。2019 年国外图像视频分享社交平台 Instagram 就推出了这类友好发言的提示功能。2022 年 2 月图像视频分享社交平台抖音也成为国内首个推出“发文警示”功能的平台。此外，抖音在预防网络暴力系统中还新增了一项功能，即“心情暖宝宝”平台助手。用户在多次违规发布评论和私信后，人工智能就会自动触发“心情暖宝宝”，以在线沟通的方式引导用户去寻求心理辅助、就诊和紧急帮助（见图 3）。诸如此类的还有知乎平台的“瓦力”机器人助手。



图 3 抖音平台的友好评论提醒、心情暖宝宝助手

5. 人工智能技术治理网络暴力的公共治理创新

21 世纪是人工智能迅速崛起的时代，在 5G、物联网、大数据、云计算等技术的突破后，全球互联网之间的交互日益紧密，为人工智能技术的发展提供了良好的内外部环境。人工智能技术在社会实际问题中的应用，如网络暴力的治理，给公共治理创新带来了挑战，同样为公共治理模式带来了创新的机遇。

5.1 向收缩性政府的转变

政府规模，是指政府的机构、职能、人员等各种要素构成的一个有机整体。政府规模不是越小越好，

也非越大越好。政府规模过小容易导致政府失灵与市场失灵，政府规模过大则会导致机构臃肿和财务负担加重，不利于公民幸福感和政府公信力的提升。强调政府治理能力的过程中，构建适度规模的政府也是当代公共治理模式转变的基本取向。人工智能的发展为处理海量公共治理数据提供了更为高效便捷的方式和手段，通过机器的自主学习和精准算法模型，人工智能可以在不受人为主观因素的干扰下，实现对海量数据更科学地整理和分析，进而为更优质的治理方案提供决策支持。以网络暴力治理为例，人工智能不仅可以让治理主体从简单且重复性的敏感词过滤和删除劳动中解放出来，既减少了人力成本，同时也有利于推进治理过程的扁平化和网络化。[11]

5.2 向精准化治理的转变

结合上文中人工智能技术治理网络暴力的实践应用可知，在大数据技术的支持下，人工智能在自主学习和创造后，能够精准了解网民的情绪变化及个人偏好，能够针对性地对每个网民建立完整的后台数据档案，并适时地提供相应的服务和满足网民的需求，为网络暴力治理提供了良好的预防机制。在公共治理的其他领域，人工智能依然能为公共服务和治理的精准化、高效率化提供了技术支撑。

5.3 加速公共治理协作方式转变

公共治理需要公共部门内部横向之间的协同合作。与此同时，与外部社会组织的共同合作也十分关键。治理网络暴力不仅是互联网平台的责任，相关公共部门如中央网信办也肩负着监管之责。只有在内外部双方协同合作的基础上才能从根本上改善和净化网络环境。利用人工智能的自动化与智能化功能获取各类公共治理所需的社会信息，可以解决中央与地方、地方与地方、政府与社会间信息不对称导致的合作难题。打造一体化和集约化的智能中心，可以在解决“数据孤岛”问题上加强部门横向和纵向的合作。

6. 总结

从厘清网络暴力的概念、成因和危害，再结合对人工智能的概述，本文试析了人工智能精准化治理网络暴力的可行性及其对公共治理的创新。人工智能技术的应用不仅可以抑制网络暴力的初起，精准化打击网络暴力行为，也可以加强对网络暴力的预防同时引导互联网良性可持续的发展。当然，网络暴力属于综合性的社会问题，如果仅仅依靠人工智能类的技术去解决网络暴力相关的问题是远远不够的。除此之外，还需要不断的完善法律法规和制定更加严格的监管措施，才能从根本上解决网络暴力问题。尽管我国在人工智能技术方面的发展日新月异，但就人工智能技术治理网络暴力而言仍然任重道远。

注释：

①数据来源：2022年8月23日，以中国知网期刊数据库（CNKI）为来源，检索流程如下：选择文献来源为“中文总库”，按“网络暴力”进行主题检索，再对“学科”一栏的文献数据量进行表格化处理。

②中国互联网络信息中心，《2021年全国未成年人互联网使用情况研究报告》，http://www.cnnic.cn/n4/2022/1201/c116-10690.html.2022-12-01/2022-12-05.

③中国互联网络信息中心，第50次《中国互联网络发展状况统计报告》，http://www.cnnic.cn/n4/2022/0914/c88-10226.html.2022-09-14/2022-12-05.

④中国互联网络信息中心，第47次《中国互联网络发展状况统计报告》，http://www.cac.gov.cn/2021-02/03/c_1613923423079314.htm.2021-02-03/2022-12-05.

⑤刘学洲事件——从小被亲生父母卖掉的刘学洲，于2021年找到了亲生父母，由于尚未成年经济无法独立，养父母也均已去世，居无定所，便提出了和亲生父母一同生活或者为他提供住所的想法，遭到了亲生父母的拒绝。随后，网络有人攻击刘学洲贪财，对他的长相、声音各个方面进行口诛笔伐。最终2022年1月刘学洲不堪网暴的压力在三亚海边吞药自杀身亡。

⑥王力宏离婚事件——2021年12月中旬，李靓蕾发表了一篇她与王力宏离婚内幕的博文，李靓蕾随后得到了网民的支持。但是后面又发了一条博文提及徐姓女歌手，因为表述含糊不清，王力宏也没能及时解释，导致徐姓女歌手被网暴，广告合约减少，收入大幅缩水。而后某陈姓作家也因为质疑李靓蕾遭到了大量的网暴。

⑦电视剧虐猫事件——某于姓导演的电视剧《当家主母》中一段猫咪中毒死亡的片段被网民怀疑是真实存在的，于是对电视剧的剧组及其导演进行了网暴，剧组多次澄清不属实，但是依然受到了网民的质疑。最后事态愈演愈烈，超六万人给电视剧评低分，导致收视率不佳。剧组无奈选择报警后，三名煽动网暴的始作俑者均被捕。

⑧德勤《人工智能产业白皮书》，https://www2.deloitte.com/cn/zh/pages/innovation/articles/china-ai-industry-

（下转第87页）